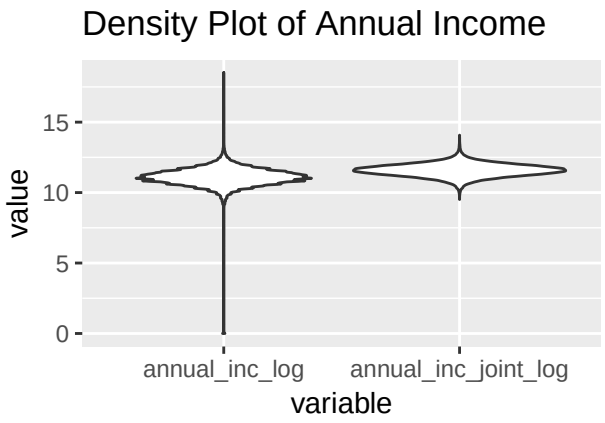
R Notebook

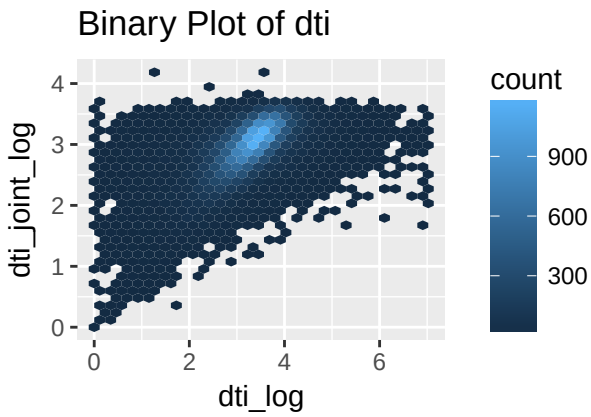
Data Preprossessing

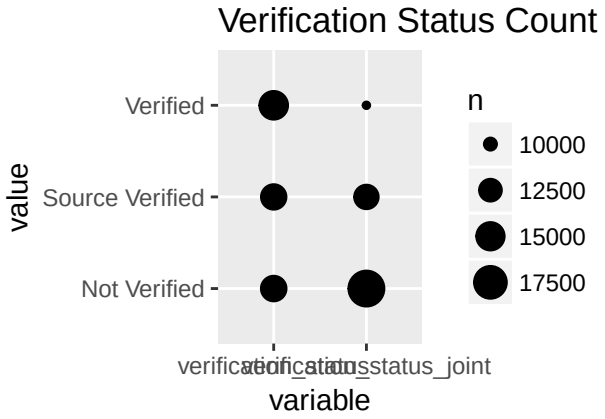
First Plot

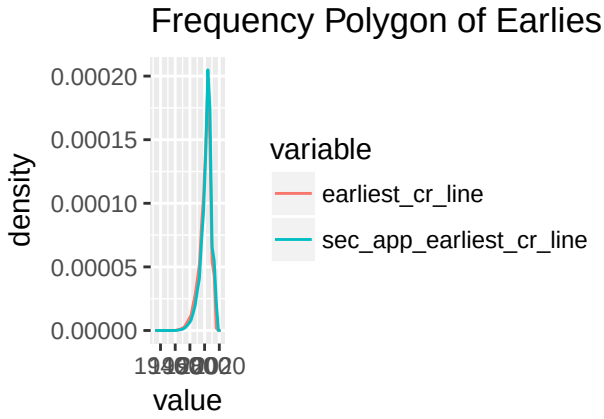
```
## .1
jointPos <- grep("_joint", names(ls2017))
ls2017.joint <- ls2017[, names(ls2017)[c(jointPos, jointPos - 1)], with = F]
setcolorder(ls2017.joint, order(names(ls2017.joint)))
# str(ls2017.joint)

# violin default already scaled:
## Income and Revolving Balance has a long tail, which is believe to be outliers (false data)
## One possible method is to use log(x+1) transform
ls2017.joint[, paste0(names(ls2017.joint)[1:6], "_log") := log(.SD + 1), .SD = names(ls2017.joint)[1:6]]


p11 <- ggplot(melt(ls2017.joint[,9:10], measure.vars = names(ls2017.joint[,9:10])),
              aes(x = variable, y = value)) + geom_violin() +
  ggtitle("Density Plot of Annual Income")


p12 <- ggplot(ls2017.joint, aes(x = dti_log, y = dti_joint_log)) + geom_hex() +
  ggtitle("Binary Plot of dti")

p13 <- ggplot(melt(na.omit(ls2017.joint[,7:8]), measure.vars = names(ls2017.joint[,7:8])),
              aes(x = variable, y = value)) + geom_count() + ggtitle("Verification Status Count")

## .2
secAppColnames <- grep("sec_app_", names(ls2017), value = T)
ls2017.secApp <- ls2017[, c(gsub("sec_app_", "", secAppColnames), secAppColnames), with = F]

p14 <- ggplot(melt(na.omit(ls2017.secApp[, c(3,13)]), measure.vars = names(ls2017.secApp)[c(3,13)]),
              aes(x = value, y = ..density.., col = variable)) + geom_freqpoly() +
  ggtitle("Frequency Polygon of Earlies Credit Line Date")

grid.arrange(p11, p12, p13, p14, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



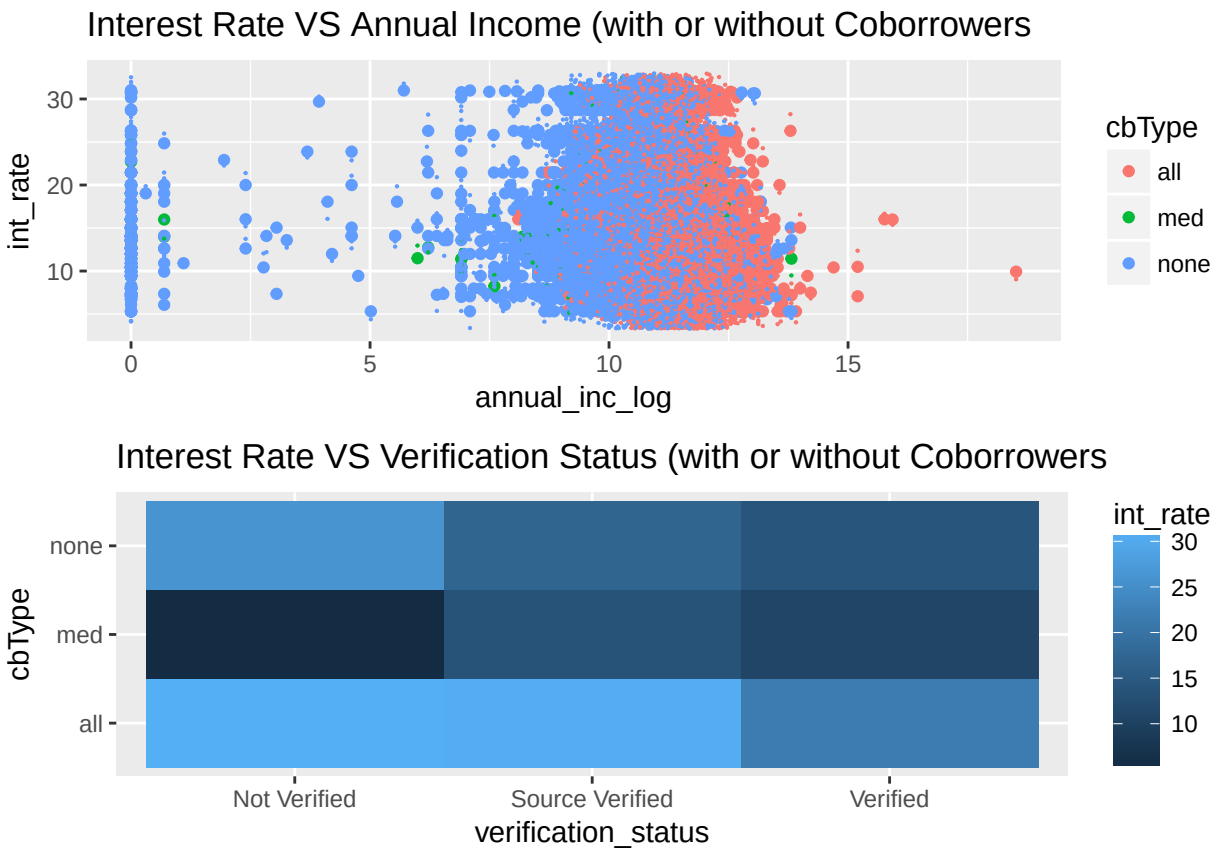






For the features with info both of borrower, one can see the similarity. The coborrower's income is higher, which might be the case parents are the co-borrower of the students. The meaning of dti is a ratio of debt payment and income. Co-borrower has a low ratio. They don't have a huge burden. The verification status is somehow interesting. Less co-borrowers are verified might due to the verification mechanism of Lending Club. Second applicants built a late credit line. They are elder. So they contact credit products later.

Second Plot



The cbType, referring to co-borrower's type, is defined as follows. All means the applicant has a co-borrower, none means no, and med denotes the applicant's co-borrower information is partially missing. If we add one response, e.g. interest rate, one can see a separate pattern against annual income under the cbType. As for verification status, the trends differs a lot among these 3 cbTypes.

A casual linear regression is also applied and result in a slight decrease of the prediction error.